



Academic performance and behavioral patterns

Kassarnig, Valentin; Mones, Enys; Bjerre-Nielsen, Andreas; Sapiezynski, Piotr; Dreyer Lassen, David; Lehmann, Sune

Published in:
E P J Data Science

Link to article, DOI:
[10.1140/epjds/s13688-018-0138-8](https://doi.org/10.1140/epjds/s13688-018-0138-8)

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

[Link back to DTU Orbit](#)

Citation (APA):
Kassarnig, V., Mones, E., Bjerre-Nielsen, A., Sapiezynski, P., Dreyer Lassen, D., & Lehmann, S. (2018). Academic performance and behavioral patterns. *E P J Data Science*, 7(1), 1-16.
<https://doi.org/10.1140/epjds/s13688-018-0138-8>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.



Academic performance and behavioral patterns

Valentin Kassarnig^{1*} , Enys Mones², Andreas Bjerre-Nielsen^{3,4}, Piotr Sapiezynski^{2,5}, David Dreyer Lassen^{3,4} and Sune Lehmann^{2,4,6*}

*Correspondence:

kassarnig@ist.tugraz.at; sljo@dtu.dk

¹Institute of Software Technology, Graz University of Technology, Graz, Austria

²Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark
Full list of author information is available at the end of the article

Abstract

Identifying the factors that influence academic performance is an essential part of educational research. Previous studies have documented the importance of personality traits, class attendance, and social network structure. Because most of these analyses were based on a single behavioral aspect and/or small sample sizes, there is currently no quantification of the interplay of these factors. Here, we study the academic performance among a cohort of 538 undergraduate students forming a single, densely connected social network. Our work is based on data collected using smartphones, which the students used as their primary phones for two years. The availability of multi-channel data from a single population allows us to directly compare the explanatory power of individual and social characteristics. We find that the most informative indicators of performance are based on social ties and that network indicators result in better model performance than individual characteristics (including both personality and class attendance). We confirm earlier findings that class attendance is the most important predictor among individual characteristics. Finally, our results suggest the presence of strong homophily and/or peer effects among university students.

Keywords: Academic performance; Data collection; Homophily; Peer effect

1 Introduction

Since research on academic achievement began to emerge as a field in the 1960s, it has guided educational policies on admissions and dropout prevention [1]. Although much of the literature has focused on higher education, the knowledge obtained on behavioral phenomena observed in colleges and universities can potentially guide research on student behavior in primary and secondary schools. A number of behavioral patterns have been linked to academic performance, such as time allocation [2], active social ties [3], sleep duration and sleep quality [4], or participation in sport activity [5]. Most of the existing studies, however, suffer from biases and limitations often associated with surveys and self-reports [6, 7], particularly when measuring social networks [8–11].

Here we investigate the performance of 538 students within a novel dataset collected as part of the *Copenhagen Network Study* (CNS), with data collection ongoing for more than two years [12]. Due to the scale of the CNS, and the inclusion of directly observed data from smartphones in place of self-reports, we are able to mitigate some of the limitations

encountered in existing ‘traditional’ studies. The strength of the CNS data is the high-resolution multi-channel measures for social interactions, including person-to-person proximity (using Bluetooth scans), calls and text messages, activity on online social networks (Facebook), and mobility traces.

The aim of our study was to better understand the impact of individual and network factors on our ability to distinguish between groups of students based on their performance. That is, we wanted to identify the ways in which low performers are significantly different from high performers and vice versa. We divide this goal into three specific objectives:

- (i) Identify individual and network factors that correlate with students’ performances.
- (ii) Analyze the importance of different sets of features for supervised learning models to classify students as low, moderate, or high performers.
- (iii) Investigate significant differences among performance groups for the most important individual and network features.

2 Related work

2.1 Individual behavior

Through a variety of methods, a large number of studies have investigated the factors that determine academic performance. Vandamme et al. [13] analyzed a broad range of individual characteristics concerning personal history, behavior, and perception. Similarly, the *StudentLife* study [14] used smartphones to collect data on student activity, social behavior, personality, and mental health. Both research groups observed correlations between performance and all feature categories, building a case that factors influencing academic performance are not limited to a single aspect of an individual’s life. Nghe et al. [15] re-framed the problem as a prediction task: using data to predict performance in a population of undergraduate and postgraduate students at two different institutions. Using a wide range of features, they predicted GPA after third year with high accuracy. One of the features included GPA after the second year; in this work we show that even without the knowledge of past achievements it is possible to explain the students’ performance levels to a large extent. Furthermore, prior research has emphasized the positive influence of attending classes [16–19]. The study by Crede et al. [19] concludes that attendance is the most accurate known predictor of academic performance; see [20] for a more detailed analysis of the impact of class attendance on academic performance based on the CNS data.

Cao et al. [21] analyzed behavioral data from the digital records of nearly 19,000 students’ smart cards, such as entering and leaving the library, having a meal in the cafeteria, or taking a shower in the dormitory. They conclude that the students’ orderliness (regularity of daily activities) is a strong predictor of academic performance. Our approach shares some similarities with [21], but the key difference is that we have investigated not only individual behavior but also the students’ social environment.

2.2 Individual traits

A large body of research at the intersection of psychology and education investigated the relationship between personality and performance, as pioneered by [22]. Many personality traits were found to be linked to academic success: Among the dimensions of the well-studied Big-Five Inventory [23] *Conscientiousness* (positive) and *Neuroticism* (negative) displayed the strongest correlation with academic performance [24–52]. The other three

dimensions showed only very weak or no correlation. Furthermore, the characteristics *Self Esteem* [53], *Satisfaction with Life* [54, 55], and *Positive Affect Schedule* [56] were also found to be positively correlated, while *Stress* [57, 58], *Depression* [59–61], and *Locus of Control* [54, 55] showed a negative effect on academic achievements.

2.3 Online social media

Only a few prior studies have investigated the impact of social media activity on academic performance, despite the growing availability of such data and undisputed presence of these media in our daily lives. The majority of existing studies found a decrease in academic performance with increasing time spent on social media [62–69]. However, not all studies confirm this result. In some studies, time spent on social media was found to be unrelated to academic performance [70, 71] or even a had positive effect on performance [72, 73].

2.4 Social interactions

There is a growing interest in the relationship between social interactions (especially on-line social interactions) and academic performance [3, 74–92]. In the relevant literature there exist two dominant approaches. The first approach focuses on the relation between own performance and that of peers [74–81], based on a hypothesis of similarity in peer achievement. The similarity between pairs of individuals connected via social ties are attributed to various aspects: selection into friendships by similarity (i.e., homophily); influence by social peers (also known as peer effect); and correlated shocks (e.g., being exposed to the same teacher). As noted by [74, 93] the issue of separating these effects is inherently difficult. The second approach emphasizes the positive influence of having a central position in the social network between students [85–90]. The majority of results in the existing research which measure social networks are, however, based on self-reports and therefore subject to various biases [8–11] that are in many ways mitigated by using smartphones to measure the social network [94]. However, it should be noted that surveys and observational studies often measure very different aspects of reality. For instance, in the case of assessing tie strengths, observational studies may be more accurate in quantifying duration and frequency variables of a relationship, while surveys can provide qualitative insights into depth and intimacy [95, 96].

3 Materials and methods

3.1 Data collection and preprocessing

Results presented in this paper are based on the data collected in the *Copenhagen Network Study* (CNS) [12]. In the CNS, dedicated smartphones were handed out to students at the Technical University of Denmark (DTU) and used as their primary phones for two years. During this period various data types were recorded: Bluetooth scans, call and text message meta data, Facebook activity logs, and mobility traces. Additionally, participating students answered a survey on personality at the beginning of the study. Due to the possibility to exit the experiment at any given point, the number of participants varied over time. We investigate the data from 538 undergraduate students for whom we have complete data.

The raw data records are cleaned and transformed to meaningful information before the analysis. Bluetooth scans are used to estimate person-to-person interactions corresponding to a physical distance of up to 10 m (30 ft) between participants. While physical

proximity is not a perfect proxy for person-to-person interactions, there is evidence that the proximity interactions are predictive of friendship in online social networks and communication using phone calls and text messages [97–99].

Facebook data was obtained via the Facebook Graph API, and contains both static friendship connections as well as various interactions on the social network. All types of interactions are treated equally. Private messages, however, are unavailable since they cannot be obtained from Facebook using the official Graph API.

The location data on the smartphones has varying accuracy depending on the providing sensor. The accuracy of the collected position can vary between a few meters for GPS locations, to hundreds of meters for cell tower location. We group the location data into 15-minute bins and use the median location of all data points with an accuracy below 80 m. In order to compute attendance we combined the smartphone locations with the person-to-person proximity obtained from Bluetooth scans. A detailed description of the method can be found in a companion paper [20].

We considered social interactions of five different channels: proximity, Facebook (friendships + interactions), calls, and text messages. For each channel we created a network to model the social relations. Note that these models are based only on the interactions among participants of the CNS. Interactions with any people outside the study were not considered. Importantly, for the proximity networks we excluded all meetings that took place during class time in order to eliminate effects caused by class co-attendance. Section B in Additional file 1 discusses further details of the creation of these network models. In the remainder of this paper, the direct neighbors in those networks are referred to as ‘peers.’

The students’ course grades were provided by DTU administration. Only courses using the Danish 7-point grading scale were considered. This scale consists of the grades 12, 10, 7, 4, 02, 00, and –3 with 12 being the best grade and 00 and –3 indicating that the student failed. The positive weighted mean grades (term or cumulative) were converted to the standard GPA scale ranging from 4.0 (best) to 0.0 (worst). Every negative mean grade was set to 0.0. Only students attending at least three courses were considered. Figure 1 illustrates the distribution of the 538 cumulative GPAs. It shows a left-skewed distribution with a mean GPA of 2.5. More information about the student population can be found in Section A of Additional file 1.

In order to increase the stability of the results we applied bootstrap resampling. Analyses were performed on 100 bootstrap samples, where each has the same size as the original sample. We report as results the mean of the bootstrap analyses with approximated standard errors described by the *Standard Error of the Mean*.

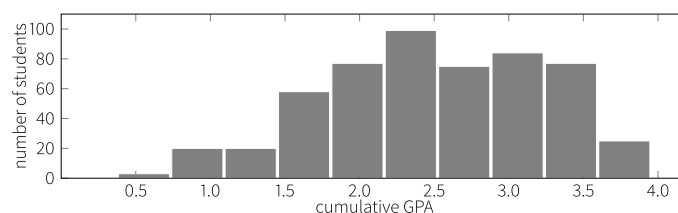


Figure 1 Distribution of cumulative GPAs. Distribution of 538 cumulative GPAs. The histogram shows a left-skewed distribution with a mean GPA of 2.5

3.2 Feature sets

To account for the different explanatory power of the individual and network aspects, we constructed four feature sets, each representing a certain aspect of life and corresponding to a specific level of information: *personality*, *individual*, *network* and *combined*.

3.2.1 Personality features

The personality features contain 16 individual personality traits obtained from questionnaires that the study participants had to fill in before receiving a phone.

3.2.2 Individual features

The individual feature set combines the 16 personality traits with behavioral and personal variables. Behavioral variables include average class attendance and the Facebook activity level (log of average number of posts per week). In terms of personal information, we added the students' gender and their study year to the feature set. Information about the sociological background of the students was not available to us.

3.2.3 Network features

For the network features we consider metrics from five different networks, each based on a different channel (texts, calls, proximity, Facebook interactions, and Facebook friendships). Despite the large number of possible features to extract from networks, we considered only the metrics that follow the main approaches found in the literature, such as the mean GPA of peers, centrality, and the fraction of low and high performing peers. However, further aspects, such as deviation, skewness, or entropy of peers' GPAs, would undoubtedly be interesting for future investigations.

The structure of the interaction networks provide further insight into how students' position in their social environment is correlated with performance. Therefore, we evaluated different centrality measures.^a Overall, the degree centrality displayed the strongest correlation and was therefore used as feature in our analyses.

3.2.4 Combined features

The combined feature set contains all 20 individual features and all 20 network features yielding a total of 40 features. See Table 1 for a complete list of features in each category. More details including descriptive statistics can be found in Section E of Additional file 1.

3.3 Approach

We use machine learning techniques to evaluate the importance of different factors on the academic performance of students. Specifically, we create supervised learning models and evaluate their performance on classifying students as low, moderate, or high performers. This framework allows us to compare our results to related work, in particular, the works by Vandamme et al. [13] and Nghe et al. [15]. Furthermore, this approach makes it easier to detect significant differences between the individual performance groups. In contrast to classical statistical modeling with test of significance, machine learning uses a hypothesis-free approach that allows us to model complex interactions driven by the data [100]. We evaluate the model performance based on the mean classification accuracy of 100 independent 10-fold cross-validations.

Table 1 Feature sets for data-driven modeling

Personality	Individual	Network	Combined
BFI: Neuroticism	Facebook activity	<i>Calls</i>	All individual features and all network features together
BFI: Openness	Class attendance	Degree Centrality	
BFI: Conscientiousness	Gender	Mean GPA of peers	
BFI: Extraversion	Study year	Fraction of low/high performing peers	
BFI: Agreeableness	+ all personality features		
Satisfaction with Life		<i>Texts</i>	
Locus of Control		Degree Centrality	
PANAS: Positive		Mean GPA of peers	
PANAS: Negative		Fraction of low/high performing peers	
Self-esteem			
Loneliness		<i>Proximity</i>	
Stress		Degree Centrality	
Depression		Mean GPA of peers	
Narcissism: Rivalry		Fraction of low/high performing peers	
Narcissism:			
Admiration		<i>FB friends</i>	
Narcissism: Overall		Degree Centrality	
		Mean GPA of peers	
		Fraction of low/high performing peers	
		<i>FB interactions</i>	
		Degree Centrality	
		Mean GPA of peers	
		Fraction of low/high performing peers	

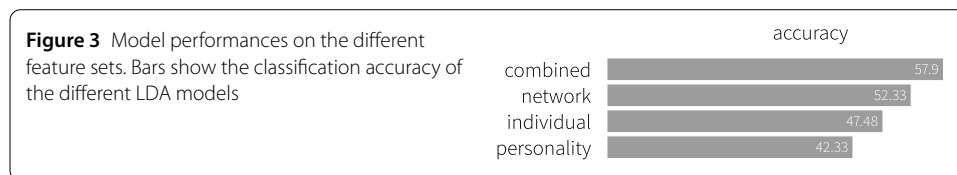
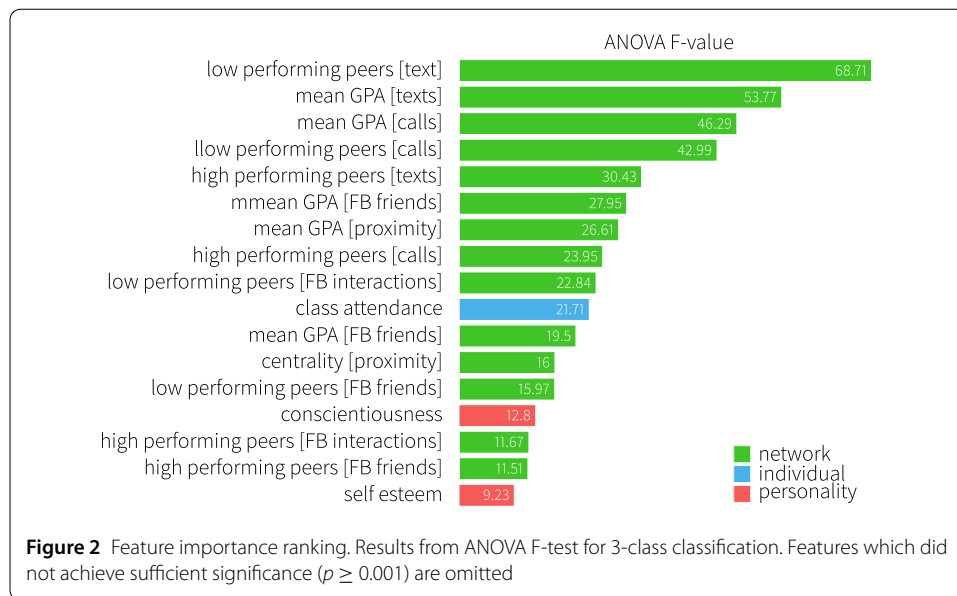
A key point to emphasize here is that while classifying students' performance levels based on current behavior might be useful in a practical context (for example to identify students in need of extra support), it is not our primary reason for using machine learning in the current study. Rather, we use machine learning as a tool for ranking and comparing features. That is, the more predictive a given feature is, the more important it is for describing performance. By training our models on features arising from many categories, previously only studied independently, we can begin to understand their relative importance, as well as their interplay in terms of academic performance.

4 Results

The following results are reported in three stages. First, we perform an ANOVA F-test on all features to identify the most important features for dividing students into performance groups. Then we utilize supervised learning models to investigate the importance and interplay of the different feature categories. Based on the results of the first two stages, we then conduct an in-depth analysis of the most expressive impact factors of each category. Our primary focus is on the social behavioral features which have only been considered to a limited extent in previous studies.

4.1 Analysis of variance

Figure 2 shows the feature importance for features achieving significance of $p < 0.001$ obtained from an ANOVA F-test.^b Although all feature categories are correlated with academic performance, the result indicates that features which describe the social networks of students have the highest explanatory power. In general, network properties dominate the results with more than half of the significant features corresponding to this category.



A potential explanation for the high impact of social relations is that the network connections may act as a proxy for previous performance, since the network features include information on the grades of others. The fraction of low performing peers as well as the mean GPA of peers contacted over text messages and calls display the highest explanatory power.^c Class attendance proves to be the most important individual feature and moreover, overall the most important one if we had no information on anyone's grades. Centrality in the proximity network is also found to be a significant descriptor with moderate importance. Among personality traits, only self-esteem and conscientiousness have significant explanatory power.

4.2 Supervised learning

In order to better understand the importance and interplay of different factors on the academic performance we utilized supervised learning techniques. We created models based on the different feature sets to classify the students as low, moderate, and high performers according to their GPAs. Each of those three groups contains the same number of students, corresponding to a baseline accuracy of 33.33%.

We use Linear Discriminant Analysis (LDA) to find an optimal model that separates the three performance classes. Figure 3 illustrates the mean results of 100 independent 10-fold cross-validations. The results show that the LDA model solely based on personality features exceeds the baseline performance by about 9 pps. Adding the four additional individual features (behavior + background info) improves the model's performance by further 5.2 pps. Using network features instead of individual features results in a performance of about 19 pps above baseline. Combining individual and network features yields a

Figure 4 Precision-recall curve. Dots represent the model performance in the low (red), moderate (green) and high (blue) performer classes. Dashed lines mark the profile of constant F_1 corresponding to the measured values for the specific class

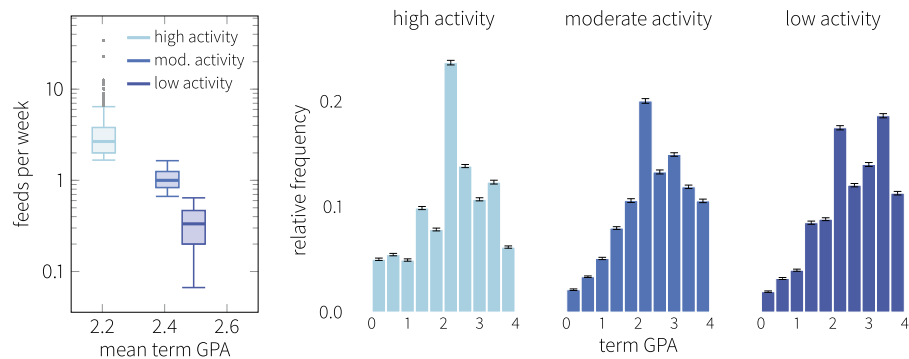
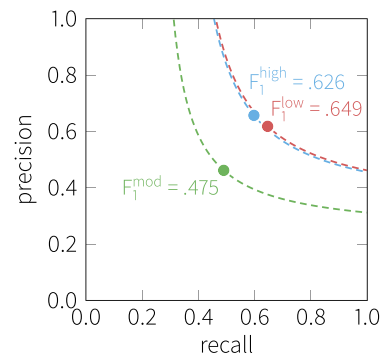


Figure 5 Facebook usage and performance in the tertiles. **(a)** Division of students into three groups of equal size according to their active Facebook updates. Each box represents a single tertile, width corresponds to the span of Facebook activity in the specific group and the x-position shows the mean term GPA. **(b)** Grade distribution inside each Facebook activity class

superior model with about 57.9% accuracy; roughly 25 pps above baseline. Figure 4 shows its achieved in-class precision and recall values along with the corresponding F_1 values. As the results indicate, once the GPA class is provided, the model has high predictive power among the low and high performers (compared to that of the moderate performers) with F_1 values of 0.649 and 0.626, respectively.

4.3 Feature analysis

4.3.1 Individual behavior

Among the considered individual effects, class attendance was found to have the highest impact on academic performance. A correlation coefficient of $r_S = 0.294$ for cumulative GPAs was determined ($p < 0.001$). An in-depth analysis of the observed class attendance patterns along with a detailed description of the method to measure attendance in the CNS dataset is discussed in [20].

The Facebook activity level measures the average number of published posts. Since the activity levels change significantly over time we consider each semester separately and use the corresponding term GPAs as measure for academic performance. This gives us up to four data points per student (one for each semester of the data collection period) for this analysis. In Fig. 5 students are divided into three groups of equal size according to their activity levels. As Fig. 5(a) shows, the distribution of posts among students is heavy-tailed

and is described by the vast majority of the students having less than 3 posts in a typical week. The distribution of term GPA values in the different tertiles reveals that, on average, students with lower activity perform better (see Fig. 5(b)). To statistically evaluate the variation in the distribution over the different tertiles, we performed a Kruskal–Wallis H-test. This test rejected the global null hypothesis with $p < 0.001$ that the medians of the groups are all equal. A follow-up Dunn multiple comparison test with Bonferroni correction revealed pair-wise differences among the tertiles: all pairs are significantly different from each other ($p < 0.001$). Thus, groups with different levels of Facebook activity have significantly different academic performances.

4.3.2 Social interactions

Based on the results presented in Fig. 2 and Fig. 3 we conclude that a student's performance can be accurately inferred from the achievements of their peers. This effect was consistently observed across different communication and interaction channels, as shown in Fig. 6. There, each channel is represented by a separate line illustrating the mean correlation of the members of each performance group and their respective peers. We can observe that regardless of the channel considered, each curve shows a strong increasing trend. This is further quantified in Table 2 which displays the corresponding correlation coefficients on the individual level. The most pronounced effect is observed for calls and text messages, which are considered to be proxies for strong social ties because this type of connection requires effort to initiate and maintain [101].

Interestingly, these channels are not dominant in the case of centrality measures. Here, proximity interactions displayed the strongest correlation among all channels. However, we found weak to moderate positive correlations in all social networks, in agreement with the existing literature [85–90].

We further assessed the validity of pairwise similarity in the network by focusing exclusively on social ties based on text messages. Figure 7 shows a scatter plot of the correlation

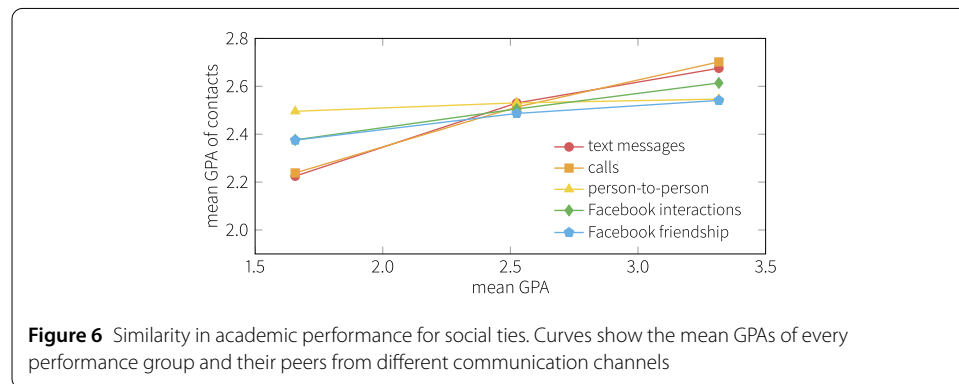


Table 2 Correlation between the cumulative GPA of the students and the mean cumulative GPA of their peers based on different communication channels. Corresponding p -values are below 0.001

Channel	r_s
Texts	0.432
Calls	0.415
Facebook interactions	0.323
Facebook friendships	0.300
Proximity	0.299

Figure 7 Correlation between performance of strong peers. For each student, we show their cumulative GPA versus the mean GPA of their peers obtained by their text messages. Color denotes density of points in arbitrary units

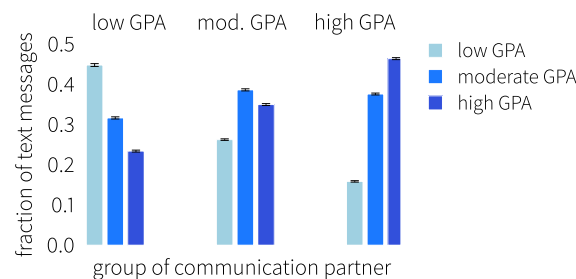
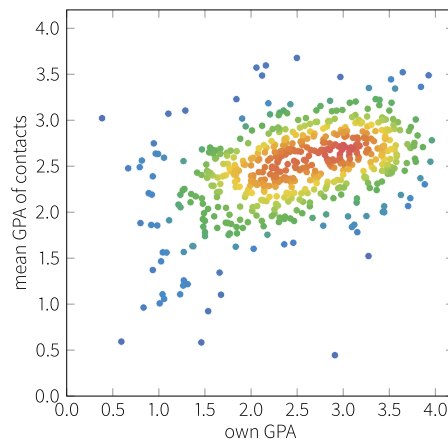


Figure 8 Own academic performance and peers' academic performance. Each histogram displays how students distribute their text messages exchanged with others over the various performance groups. Groups are defined by tertiles based on their cumulative GPA

between the own GPA and mean GPA of the texting peers for every student in the dataset. Once again, we observe a clear linear trend; the trend is especially strong in the region where the majority of the students is located (GPAs in the range between 2 and 3). In Fig. 8 we divided the population into tertiles based on the GPA and calculated the fraction of text messages exchanged with members of the different groups. Beyond the correlation, we can see that the students' communication in each group is dominated by members of the same group. This observation further underlines the importance of the social environment for academic success.

5 Discussion

For the participants of the CNS, we found that the peers' academic performance has a strong explanatory power for academic performance of individuals. We observed this effect across different channels of social interactions with calls and text messages showing the strongest correlations, further emphasizing the phenomena. As mentioned in the literature review, this effect could be caused by either peer effects (adaption) or homophily (selection). It should be noted that GPA information is used here as target and, in aggregated form, also as network feature. This allows us to analyze and understand the relationships among peers; but should be taken into account when framing the problem as prediction task.

We found network centrality to have a positive correlation with academic performance, in agreement with the literature [85–90]. However, among all types of interaction networks, only proximity networks exhibited a strong effect. A possible limitation in measuring centrality is that the mere physical proximity of two individuals does not necessarily involve direct communication. Nevertheless, it is reasonable to expect an increased level of information exchange in a group of individuals if they are in close proximity, which was the case in our dataset.^d

Consistent with findings in existing literature, we found that class attendance showed the strongest correlation with academic performance when we consider only individual effects [16, 18, 19, 102–106]. We also found that Facebook activity has a negative relation to academic performance—also in agreement with the majority of the studies that investigated Facebook and social media usage [62–69]. We note, however, that our data is limited to Facebook activities such as posting a status update or uploading a picture etc, and that we have no information regarding ‘passive’ Facebook usage, such as scrolling and reading. Also, our data does not include direct messages which may constitute a relevant fraction of communications performed via the social network site.

The analysis of the different personality traits revealed that two characteristics, namely conscientiousness and self-esteem, have considerable explanatory power for academic success. These two traits reached a correlation coefficient between 0.2 and 0.3 corresponding to the upper limit achievable for any correlation with a personality trait, according to Mischel [107]. The impact of other investigated characteristics could not be confirmed with proper significance. These results agree with existing literature [24–53].

In the supervised learning experiment we achieved a classification accuracy of around 25 percentage points above baseline, a result similar to that of Vandamme et al. [13]. While the classification accuracy is similar, comparing our results with theirs is difficult because of the very different feature sets and experimental setups. Vandamme et al. [13] use nearly ten times as many features to build a model as we did. In addition, the accuracy of Vandamme et al. [13] is driven by using prior achievement (grades), which is known to be a strong predictor of performance (e.g. due to persistence of skill and motivation). We note here that a potential reason for the similarity in performance to Vandamme et al. [13] could be that the network features used in our study include the grades of others in the network. Thus, if the network homophily with respect to academic performance is sufficiently strong, the average performance of others could serve as a proxy for each individual’s academic achievements.

Networks originating from different channels were treated separately because each network provides different information. For future studies it could be interesting to combine them and create multiplex network models which capture interactions across multiple channels and provide more information about the actual tie strength.

In summary, our findings—together with the results in the literature—emphasize that there is a considerable dependence of academic performance on personality and social environment. This experiment is by no means an attempt to be exhaustive of the possibilities for impact factors. Rather, we hope that this demonstration will stir interest to further study the impact of the social environment on academic success, as well as the interplay of individual and network factors.

5.1 Limitations

Although we utilized wider and more detailed data than most other studies, our approach also has important limitations which need to be taken into account. First, we only observed students from a single, technical, Danish university. For this reason, the findings may not be generalizable to students at other institutions, of other academic disciplines or with other demographics. Furthermore, only a subset of all the students at DTU participated in our study—for first year students the rate was around 40%. Although we observed a high degree of variation with respect to behavioral and network measures as well as academic performance, our sample may not be representative of the whole student population. Our measures of ego-networks and model estimates reflect only the smaller (and not closed) community of students in the CNS within the larger population of students.

Although direct measures overcome a lot of the limitations of surveys and self-reports, they continue to be affected by standard concerns over observational data, including selection bias, information bias, and confounding [108]. In particular, confounding plays a big role in our study as there are many factors that we were unable to capture but provenly affect the academic performance directly or interplay with other observed factors. For instance, many socio-economic variables have been identified as good predictors for academic achievements [109–112] but unfortunately such data was not available to us. There was also some tendency of selection into the study as the average student in the study tends to achieve higher grades than non-participants [113]. Furthermore, investigations on the CNS data have revealed, that findings differ slightly for men and women [114].

Social network observations were limited to phone calls/texts, meetings, and Facebook activities. Although these are arguably some of the most important means of communication, some students may communicate via other smartphone apps. Our method of inferring attendance is also subject to some noise (as thoroughly discussed in [20]). Furthermore, it does not imply in-class participation nor attention to the taught material.

Although we have identified many factors that correlate with academic performance, we make no claims regarding causality. The question of establishing causality from purely observational data is far from trivial. Thus, while being beyond the scope of this work we consider this question as promising and interesting for future research.

Additional material

Additional file 1: Supporting information. (PDF 181 kB)

Acknowledgements

Due to privacy implications we cannot share data but researchers are welcome to visit and work under our supervision.

Funding

This work was supported by the Villum Foundation, the Danish Council for Independent Research, University of Copenhagen (via the UCPH-2016 grant Social Fabric and The Center for Social Data Science) and Economic Policy Research Network (EPRN).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors contributed equally to this work. All authors read and approved the final manuscript.

Author details

¹Institute of Software Technology, Graz University of Technology, Graz, Austria. ²Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark. ³Department of Economics, University of

Copenhagen, Copenhagen, Denmark. ⁴Center for Social Data Science, University of Copenhagen, Copenhagen, Denmark. ⁵College of Information and Computer Science, Northeastern University, Boston, USA. ⁶The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark.

Endnotes

- ^a Details on the evaluation can be found in Section C of Additional file 1.
- ^b Note that F-test should not be interpreted literally here, as the assumption of identical independent draws of errors is likely to be violated due to correlation of errors in the network. Rather, we use it only as a guide to select features.
- ^c The reliability of this observation has been validated by a permutation test—see Section D of Additional file 1.
- ^d The CNS uses (thresholded) Bluetooth visibility as an indicator of person-to-person proximity.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 17 August 2017 Accepted: 17 April 2018 Published online: 24 April 2018

References

1. Lavin DE (1965) The prediction of academic performance. Russell Sage Foundation, New York
2. Macan TH, Shahani C, Dipboye RL, Phillips AP (1990) College students' time management: correlations with academic performance and stress. *J Educ Psychol* 82(4):760
3. Gašević D, Zouaq A, Janzen R (2013) "Choose your classmates, your GPA is at stake!" The association of cross-class social ties and academic performance. *Am Behav Sci* 57(10):1460–1479
4. Curcio G, Ferrara M, De Gennaro L (2006) Sleep loss, learning capacity and academic performance. *Sleep Med Rev* 10(5):323–337
5. Singh A, Uijtendewilligen L, Twisk JW, Van Mechelen W, Chinapaw MJ (2012) Physical activity and performance at school: a systematic review of the literature including a methodological quality assessment. *Arch Pediatr Adolesc Med* 166(1):49–55
6. Van de Mortel TF et al (2008) Faking it: social desirability response bias in self-report research. *Aust J Adv Nurs* 25(4):40
7. Junco R (2013) Comparing actual and self-reported measures of Facebook use. *Comput Hum Behav* 29(3):626–631
8. Kumbasar E, Rommey AK, Batchelder WH (1994) Systematic biases in social perception. *Am J Sociol* 100(2):477–505
9. O'Connor KM, Gladstone E (2015) How social exclusion distorts social network perceptions. *Soc Netw* 40:123–128
10. Freeman LC (1992) Filling in the blanks: a theory of cognitive categories and the structure of social affiliation. *Soc Psychol Q* 55:118–127
11. Bernard HR, Killworth P, Kronenfeld D, Sailer L (1984) The problem of informant accuracy: the validity of retrospective data. *Annu Rev Anthropol* 13(1):495–517
12. Stopczynski A, Sekara V, Sapiezynski P, Cuttone A, Madsen MM, Larsen JE, Lehmann S (2014) Measuring large-scale social networks with high resolution. *PLoS ONE* 9(4):95978
13. Vandamme J-P, Meskens N, Superby J-F (2007) Predicting academic performance by data mining methods. *Educ Econ* 15(4):405–419
14. Wang R, Chen F, Chen Z, Li T, Harari G, Tignor S, Zhou X, Ben-Zeev D, Campbell AT (2014) Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In: Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing. ACM, New York, pp 3–14
15. Nghe NT, Janeczek P, Haddawy P (2007) A comparative analysis of techniques for predicting academic performance. In: Frontiers in education conference-global engineering: knowledge without borders, opportunities without passports, 2007. FIE '07. 37th annual. IEEE, New York, pp 2–7
16. Buckalew L, Daly JD, Coffield K (1986) Relationship of initial class attendance and seating location to academic performance in psychology classes. *Bull Psychon Soc* 24(1):63–64
17. Marburger DR (2006) Does mandatory attendance improve student performance? *J Econ Educ* 37(2):148–155
18. Chen J, Lin T-F (2008) Class attendance and exam performance: a randomized experiment. *J Econ Educ* 39(3):213–227
19. Credé M, Roch SG, Kieszczyńska UM (2010) Class attendance in college: a meta-analytic review of the relationship of class attendance with grades and student characteristics. *Rev Educ Res* 80(2):272–295
20. Kassarnig V, Bjerre-Nielsen A, Mones E, Lehmann S, Lassen DD (2017) Class attendance, peer similarity, and academic performance in a large field study. *PLoS ONE* 12(11):e0187078. <https://doi.org/10.1371/journal.pone.0187078>
21. Cao Y, Lian D, Rong Z, Shi J, Wang Q, Wu Y, Yao H, Zhou T (2017) Orderliness predicts academic performance: Behavioral analysis on campus lifestyle. arXiv preprint. arXiv:1704.04103
22. Prociuk TJ, Breen LJ (1974) Locus of control, study habits and attitudes, and college academic performance. *J Psychol* 88(1):91–95
23. Goldberg LR (1993) The structure of phenotypic personality traits. *Am Psychol* 48(1):26
24. Dollinger SJ, Orf LA (1991) Personality and performance in "personality": conscientiousness and openness. *J Res Pers* 25(3):276–284
25. Goff M, Ackerman PL (1992) Personality-intelligence relations: assessment of typical intellectual engagement. *J Educ Psychol* 84(4):537
26. Rothstein MG, Paunonen SV, Rush JC, King GA (1994) Personality and cognitive ability predictors of performance in graduate business school. *J Educ Psychol* 86(4):516
27. Wolfe RN, Johnson SD (1995) Personality as a predictor of college performance. *Educ Psychol Meas* 55(2):177–185
28. De Fruyt F, Mervielde I (1996) Personality and interests as predictors of educational streaming and achievement. *Eur J Pers* 10(5):405–425

29. Paunonen SV (1998) Hierarchical organization of personality and prediction of behavior. *J Pers Soc Psychol* 74(2):538
30. Busato VV, Prins FJ, Elshout JJ, Hamaker C (2000) Intellectual ability, learning style, personality, achievement motivation and academic success of psychology students in higher education. *Pers Individ Differ* 29(6):1057–1068
31. Paunonen SV, Ashton MC (2001) Big five predictors of academic achievement. *J Res Pers* 35(1):78–90
32. Gray EK, Watson D (2002) General and specific traits of personality and their relation to sleep and academic performance. *J Pers* 70(2):177–206
33. Lievens F, Coetsier P, De Fruyt F, De Maeseneer J (2002) Medical students' personality characteristics and academic performance: a five-factor model perspective. *Med Educ* 36(11):1050–1056
34. Bauer KW, Liang Q (2003) The effect of personality and precollege characteristics on first-year activities and academic performance. *J Coll Stud Dev* 44(3):277–290
35. Chamorro-Premuzic T, Furnham A (2003) Personality traits and academic examination performance. *Eur J Pers* 17(3):237–250
36. Chamorro-Premuzic T, Furnham A (2003) Personality predicts academic performance: evidence from two longitudinal university samples. *J Res Pers* 37(4):319–338
37. Diseth Å (2003) Personality and approaches to learning as predictors of academic achievement. *Eur J Pers* 17(2):143–155
38. Farsides T, Woodfield R (2003) Individual differences and undergraduate academic success: the roles of personality, intelligence, and application. *Pers Individ Differ* 34(7):1225–1243
39. Furnham A, Chamorro-Premuzic T, McDougall F (2002) Personality, cognitive ability, and beliefs about intelligence as predictors of academic performance. *Learn Individ Differ* 14(1):47–64
40. Lounsbury JW, Sundstrom E, Loveland JM, Gibson LW (2003) Intelligence, "big five" personality traits, and work drive as predictors of course grade. *Pers Individ Differ* 35(6):1231–1239
41. Phillips P, Abraham C, Bond R (2003) Personality, cognition, and university students' examination performance. *Eur J Pers* 17(6):435–448
42. Duff A, Boyle E, Dunleavy K, Ferguson J (2004) The relationship between personality, approach to learning and academic performance. *Pers Individ Differ* 36(8):1907–1920
43. Furnham A, Chamorro-Premuzic T (2004) Personality and intelligence as predictors of statistics examination grades. *Pers Individ Differ* 37(5):943–955
44. Hair P, Hampson SE (2006) The role of impulsivity in predicting maladaptive behaviour among female students. *Pers Individ Differ* 40(5):943–952
45. Conard MA (2006) Aptitude is not enough: how personality and behavior predict academic performance. *J Res Pers* 40(3):339–346
46. Barchard KA (2003) Does emotional intelligence assist in the prediction of academic success? *Educ Psychol Meas* 63(5):840–858
47. Langford PH (2003) A one-minute measure of the big five? Evaluating and abridging Shafer's (1999a) big five markers. *Pers Individ Differ* 35(5):1127–1140
48. Oswald FL, Schmitt N, Kim BH, Ramsay LJ, Gillespie MA (2004) Developing a biodata measure and situational judgment inventory as predictors of college student performance. *J Appl Psychol* 89(2):187
49. Leong FT, Gibson LW, Lounsbury JW, Huffstetler BC (2005) Sense of identity and collegiate academic achievement. *J Coll Stud Dev* 46(5):501–514
50. Ridgell SD, Lounsbury JW (2004) Predicting academic success: general intelligence, "big five" personality traits, and work drive. *Coll Stud J* 38(4):607
51. Komarraju M, Karau SJ, Schmeck RR (2009) Role of the big five personality traits in predicting college students' academic motivation and achievement. *Learn Individ Differ* 19(1):47–52
52. Nofle EE, Robins RW (2007) Personality predictors of academic outcomes: big five correlates of GPA and SAT scores. *J Pers Soc Psychol* 93(1):116
53. Lane J, Lane AM, Kyprianou A (2004) Self-efficacy, self-esteem and their impact on academic performance. *Soc Behav Pers Int J* 32(3):247–256
54. Lepp A, Barkley JE, Karpinski AC (2014) The relationship between cell phone use, academic performance, anxiety, and satisfaction with life in college students. *Comput Hum Behav* 31:343–350
55. Chow HP (2005) Life satisfaction among university students in a Canadian prairie city: a multivariate analysis. *Soc Indic Res* 70(2):139–150
56. Saklofske DH, Austin EJ, Mastoras SM, Beaton L, Osborne SE (2012) Relationships of personality, affect, emotional intelligence and coping with student stress and academic success: different patterns of association for stress and success. *Learn Individ Differ* 22(2):251–257
57. Stewart SM, Lam T, Betson C, Wong C, Wong A (1999) A prospective analysis of stress and academic performance in the first two years of medical school. *Med Educ Oxf* 33(4):243–250
58. Akgun S, Ciarrochi J (2003) Learned resourcefulness moderates the relationship between academic stress and academic performance. *Educ Psychol* 23(3):287–294
59. Haines ME, Norris MP, Kashy DA (1996) The effects of depressed mood on academic performance in college students. *J Coll Stud Dev* 37(5):519–526
60. Leach J (2009) The relationship between depression and college academic performance. *Coll Stud J* 43(2):325
61. Owens M, Stevenson J, Hadwin JA, Norgate R (2012) Anxiety and depression in academic performance: an exploration of the mediating factors of worry and working memory. *Sch Psychol Int* 33(4):433–449
62. Maqableh MM, Rajab L, Quteshat W, Moh'd Taisir Masa R, Khatib T, Karajeh H et al (2015) The impact of social media networks websites usage on students' academic performance. *Commun Netw* 7(4):159
63. Al-Menayes JJ (2015) Social media use, engagement and addiction as predictors of academic performance. *Int J Psychol Stud* 7(4):86
64. Al-Menayes JJ (2014) The relationship between mobile social media use and academic performance in university students. *New Media Mass Commun* 25:23–29
65. Karpinski AC, Kirschner PA, Ozer I, Mellott JA, Ochwo P (2013) An exploration of social networking site use, multitasking, and academic performance among United States and European university students. *Comput Hum Behav* 29(3):1182–1192

66. Paul JA, Baker HM, Cochran JD (2012) Effect of online social networking on student academic performance. *Comput Hum Behav* 28(6):2117–2127
67. Junco R (2012) The relationship between frequency of Facebook use, participation in Facebook activities, and student engagement. *Comput Educ* 58(1):162–171
68. Jacobsen WC, Forste R (2011) The wired generation: academic and social outcomes of electronic media use among university students. *Cyberpsychol Behav Soc Netw* 14(5):275–280
69. Kirschner PA, Karpinski AC (2010) Facebook® and academic performance. *Comput Hum Behav* 26(6):1237–1245
70. Pasek J, Hargittai E et al (2009) Facebook and academic performance: reconciling a media sensation with data. *First Monday* 14(5)
71. Ainin S, Naqshbandi MM, Moghavi S, Jaafar NI (2015) Facebook usage, socialization and academic performance. *Comput Educ* 83:64–73
72. Kolek EA, Saunders D (2008) Online disclosure: an empirical examination of undergraduate Facebook profiles. *NASPA J* 45(1):1–25
73. Tayseer M, Zoghieb F, Alcheikh I, Awadallah MN (2014) Social network: academic & social impact on college students. Retrieved 20th November
74. Sacerdote B (2001) Peer effects with random assignment: results for dartmouth roommates. *Q J Econ* 116:681–704
75. Zimmerman DJ (2003) Peer effects in academic outcomes: evidence from a natural experiment. *Rev Econ Stat* 85(1):9–23
76. Stinebrickner R, Stinebrickner TR (2006) What can be learned about peer effects using college roommates? Evidence from new survey data and students from disadvantaged backgrounds. *J Public Econ* 90(8):1435–1454
77. Carrell SE, Sacerdote BJ, West JE (2013) From natural variation to optimal policy? The importance of endogenous peer group formation. *Econometrica* 81(3):855–882
78. Vitale MP, Porzio GC, Doreian P (2016) Examining the effect of social influence on student performance through network autocorrelation models. *J Appl Stat* 43(1):115–127
79. Smirnov I, Thurner S (2016) Formation of homophily in academic performance: students prefer to change their friends rather than performance. *arXiv preprint. arXiv:1606.09082*
80. Poldin O, Valeeva D, Yudkevich M (2013) How social ties affect peer group effects: case of university students. Higher School of Economics Research Paper No. WP BPR 15
81. Mayer A, Puller SL (2008) The old boy (and girl) network: social network formation on university campuses. *J Public Econ* 92(1):329–347
82. Yuan YC, Gay G, Hembrooke H (2006) Focused activities and the development of social capital in a distributed learning “community”. *Inf Soc* 22(1):25–39
83. Rizzuto TE, LeDoux J, Hatala JP (2009) It’s not just what you know, it’s who you know: testing a model of the relative importance of social networks to academic performance. *Soc Psychol Educ* 12(2):175–189
84. Tomás-Miquel J-V, Expósito-Langa M, Nicolau-Julíá D (2016) The influence of relationship networks on academic performance in higher education: a comparative study between students of a creative and a non-creative discipline. *High Educ* 71:307–322
85. Sparrowe RT, Liden RC, Wayne SJ, Kraimer ML (2001) Social networks and the performance of individuals and groups. *Acad Manag J* 44(2):316–325
86. Smith RA, Peterson BL (2007) “Psst... what do you think?” The relationship between advice prestige, type of advice, and academic performance. *Commun Educ* 56(3):278–291
87. Hommes J, Rienties B, De Grave W, Bos G, Schuwirth L, Scherpbier A (2012) Visualising the invisible: a network approach to reveal the informal social side of student learning. *Adv Health Sci Educ* 17(5):743–757
88. Baldwin TT, Bedell MD, Johnson JL (1997) The social fabric of a team-based mba program: network effects on student satisfaction and performance. *Acad Manag J* 40(6):1369–1397
89. Yang H, Tang J (2003) Effects of social network on students, performance: a web-based forum study in Taiwan. *J Asynchron Learn Netw* 7(3):93. Retrieved 11 26, 2011
90. Cho H, Gay G, Davidson B, Ingraffea A (2007) Social networks, communication styles, and learning performance in a CSCL community. *Comput Educ* 49(2):309–329
91. Thomas SL (2000) Ties that bind: a social network approach to understanding student integration and persistence. *J High Educ* 71:591–615
92. Johnson DW, Johnson RT (1984) Structuring groups for cooperative learning. *J Manag Educ* 9(4):8–17
93. Manski CF (1993) Identification of endogenous social effects: the reflection problem. *Rev Econ Stud* 60(3):531–542
94. Eagle N, Pentland AS, Lazer D (2008) Mobile phone data for inferring social network structure. In: *Social computing, behavioral modeling, and prediction*. Springer, Berlin, pp 79–88
95. Marsden PV, Campbell KE (1984) Measuring tie strength. *Soc Forces* 63(2):482–501
96. Newcomb AF, Bagwell CL (1995) Children’s friendship relations: a meta-analytic review. *Psychol Bull* 117(2):306
97. Sekara V, Lehmann S (2014) The strength of friendship ties in proximity sensor data. *PLoS ONE* 9(7):100915
98. Sapiezynski P, Stopczynski A, Wind DK, Leskovec J, Lehmann S (2017) Inferring person-to-person proximity using wifi signals. *Proc ACM Interact Mob Wearable Ubiquitous Technol* 1(2):24
99. Eagle N, Pentland AS, Lazer D (2009) Inferring friendship network structure by using mobile phone data. *Proc Natl Acad Sci USA* 106(36):15274–15278
100. Valletta JJ, Torney C, Kings M, Thornton A, Madden J (2017) Applications of machine learning in animal behaviour studies. *Anim Behav* 124:203–220
101. Van Cleemput K (2010) “I’ll see you on IM, text, or call you”: a social network approach of adolescents’ use of communication media. *Bull Sci Technol Soc* 30(2):75–85
102. Stanca L (2006) The effects of attendance on academic performance: panel data evidence for introductory microeconomics. *J Econ Educ* 37(3):251–266
103. Van Blerkom ML (1992) Class attendance in undergraduate courses. *J Psychol* 126(5):487–494
104. Brocato J (1989) How much does coming to class matter? Some evidence of class attendance and grade performance. *Educ Res Q* 13(3):2–6
105. Gump SE (2005) The cost of cutting class: attendance as a predictor of success. *Coll Teach* 53(1):21–26

106. Lin T-F, Chen J (2006) Cumulative class attendance and exam performance. *Appl Econ Lett* 13(14):937–942
107. Mischel W (2013) *Personality and assessment*. Lawrence Erlbaum Associates, Mahwah
108. Hill HA, Kleinbaum DG (2000) Bias in observational studies. In: *Encyclopedia of biostatistics*
109. DeBerard MS, Spielmans G, Julka D (2004) Predictors of academic achievement and retention among college freshmen: a longitudinal study. *Coll Stud J* 38(1):66–80
110. Cohn E, Cohn S, Balch DC, Bradley J (2004) Determinants of undergraduate GPAs: SAT scores, high-school GPA and high-school rank. *Econ Educ Rev* 23(6):577–586
111. White KR (1982) The relation between socioeconomic status and academic achievement. *Psychol Bull* 91(3):461
112. Sirin SR (2005) Socioeconomic status and academic achievement: a meta-analytic review of research. *Rev Educ Res* 75(3):417–453
113. Bjerre-Nielsen A, Dreyer Lassen D (2017) Opportunity and similarity in dynamic friendships. Technical report
114. Sapiezynski P, Kassarnig V, Wilson C, Lehmann S, Mislove A (2017) Academic performance prediction in a gender-imbalanced environment. In: *FATREC workshop on responsible recommendation proceedings*

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)